University of Málaga

Health Engineering

Laboratory Task

Support Vector Machines

Author

Alejandro Domínguez Recio

Course

Intelligent Systems

Teachers

Enrique Domínguez Merino

Jesús de Benito Picazo

Introduction

In this practice we going to evaluate three diferents dataset with the SMO classifier. We will evaluate the dataset with different kernel functions as well as their parameters.To evaluate which dataset is more optimal with this classifier we going to describe the performance measures seen in class like confusion matrix, acurracy, precision, fallout, recall, F-measure and area under ROC curve. Also we going to detail the characteristics of the different datasets.

How are we going to do it?

To do the different studies we will have the following structure in each of then.

- **Dataset context**

In this part we going to write a little introduction of the dataset context commenting on data type and prediction target.

- **Number of classes**

We going to describe the number of classes as if it is binary o multiclass. To do that we going to observe the atributtes weka panel and we going to select the class attribute and depending on the number of values that it takes, we will determine if it is binary or multiclass.

- **Number of attributes**

For the number of attributes we going to inspect the weka attributes panel or open the file in text mode and inspect the characteristics.

- **Number of samples or instances**

To know the number of samples we can proceed as in the previous section by inspecting the weka panels or opening the document in text mode and seeing its characteristics.

➔ ***Performance metric values by configuration***

Accuracy

We can obtain the acurracy in two ways, one of them is directly from the classifier out and the another one is by calculating the number of correctly predicted examples divided by the total number of examples.

Precision

This measurement shows the positive predictive value, higher is bettter. We are going to obtain the necessary values from the confusion matrix and perform the following calculation TP/(TP+FP).

Fallout

This measurement shows the false positive rate, lower is better. We are going to obtain the necessary values from the confusion matrix and perform the following calculation FP/(FP+TN).

Recall

This measurement shows the true positive rate, higher is better. We are going to obtain the necessary values from the confusion matrix and perform the following calculation TP/(TP+FN).

F-measure

This measurement provides a single score that balances both the concers of precision and recall in one number, higher is better. We are going to obtain the necessary values of the previous measurements, precision and recall, and perform the following calculation 2*(Precision*Recall)/(Precision+Recall).

➔ ROC curve and the area under the curve

We are going to obtain this measurement by visualizing the threshold curve in Weka. The area under the ROC curve is a number in the interval [0,1], a higher value is better. This measure show the trade-off between the ratios of false positives and false negatives. This measurement is very useful when we faced with unbalanced data.

*The previous points will be repeated for each dataset

➔ Evaluation of the results

In this section we are goint to compare the results of the different performance measures.

➔ Differences between datasets

Here we are going to describe the main differences in the data of the different datasets.

➔ Last conclusions

Finally, we are going explain possible reasons why some datasets perform better with the SMO classifier than others. We will support our conclusions on the performance measures taken.

BREAST CANCER DATASET

➔Introduction

The main idea of this classification problem is given a dataset about patients with a series of characteristics which determine if they have recurrence events or no recurrence events of breast cancer train a SMO classification model and evaluate its performance.

In this dataset we have 286 instances in total of which 201 of one class and 85 instances of another class. Each instance has ten attributes, one of them is the class attribute. Because of we have only two classes we are faced with a binary classification problem.

➔Number of classes

In this case our class attribute can take only **two *values***, *recurrence events* and *no recurrence events*. Because of that we are faced to a ***binary classification problem***.

➔Number of attributes

In this case we have **ten attributes** which are *class, age, menopause, tumor-size, inv-nodes, node-caps, deg-malig, breast, breast-quad* and *irradiat.*

➔Number of samples

In this case in particular the number of samples or instances is **286.**

➔ Performance metric values by configuration

- **Configuration PolyKernel 1**

  Exponencial = 1

TP Rate	*FP Rate*	*Precision*	*Recall*	*F-Measure*	*MCC*	*ROC*
0,866	0,671	0,753	0,866	0,806	0,226	0,598

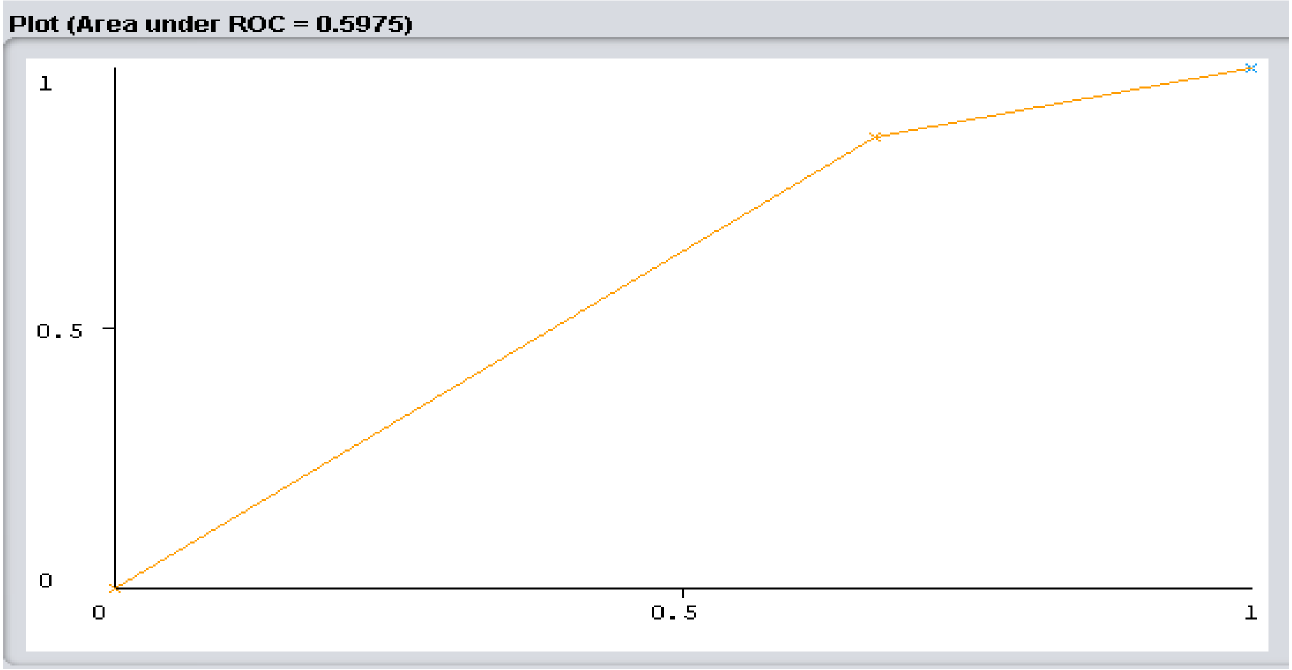


Ilustración 1: Graphic ROC Breast Cancer

- **Configuration PolyKernel 2**

  Exponencial = 10

TP Rate	*FP Rate*	*Precision*	*Recall*	*F-Measure*	*MCC*	*ROC*
0,547	0,376	0,775	0,547	0,641	0,156	0,585

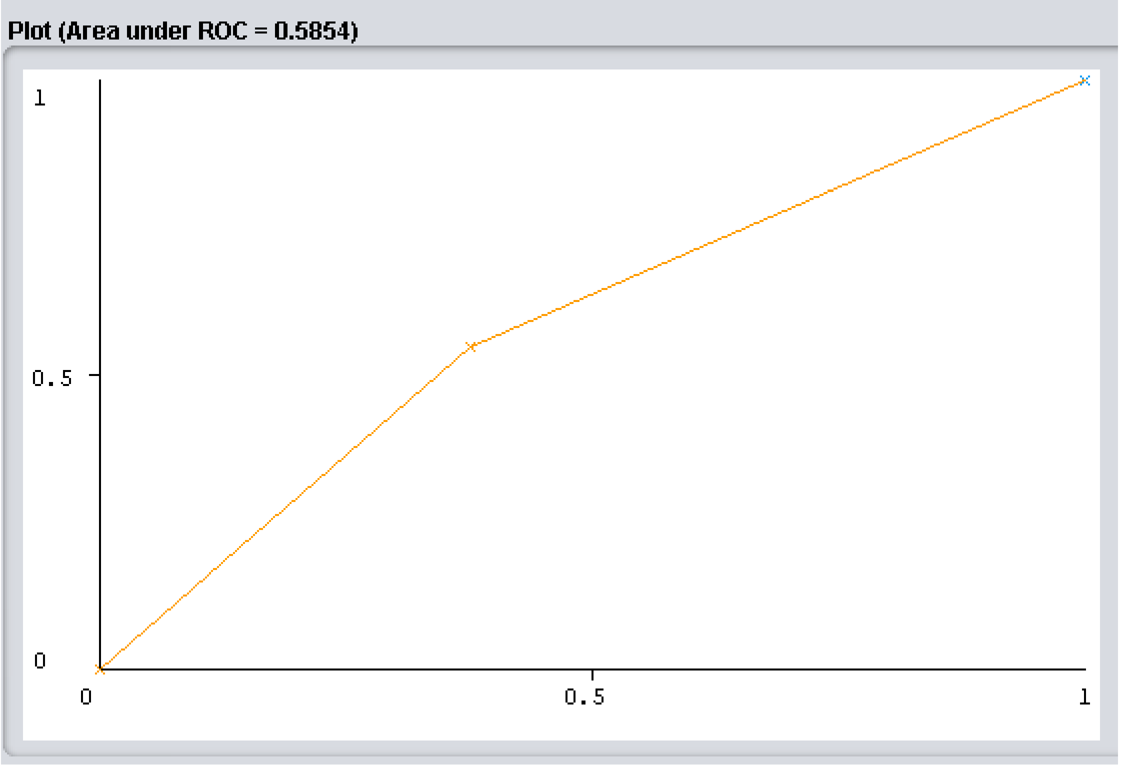


Ilustración 2: ROC BreastCancer2

- **Configuration PolyKernel 3**

  Exponencial = 100

TP Rate	*FP Rate*	*Precision*	*Recall*	*F-Measure*	*MCC*	*ROC*
1,000	1,000	0,703	1,000	0,825	?	0,500

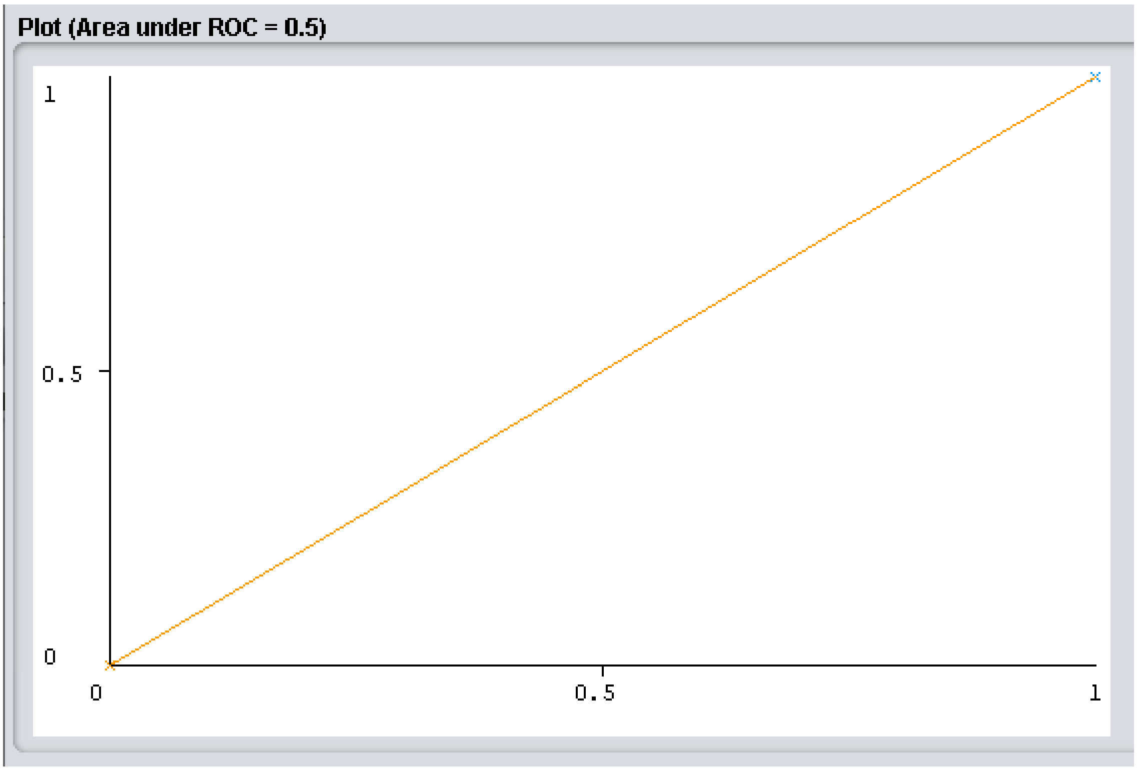


Ilustración 3: ROC BreastCancer3

- **Configuration Normalized Poly Kernel 1**

  Exponencial = 1

TP Rate	*FP Rate*	*Precision*	*Recall*	*F-Measure*	*MCC*	*ROC*
0,896	0,729	0,744	0,896	0,813	0,210	0,583

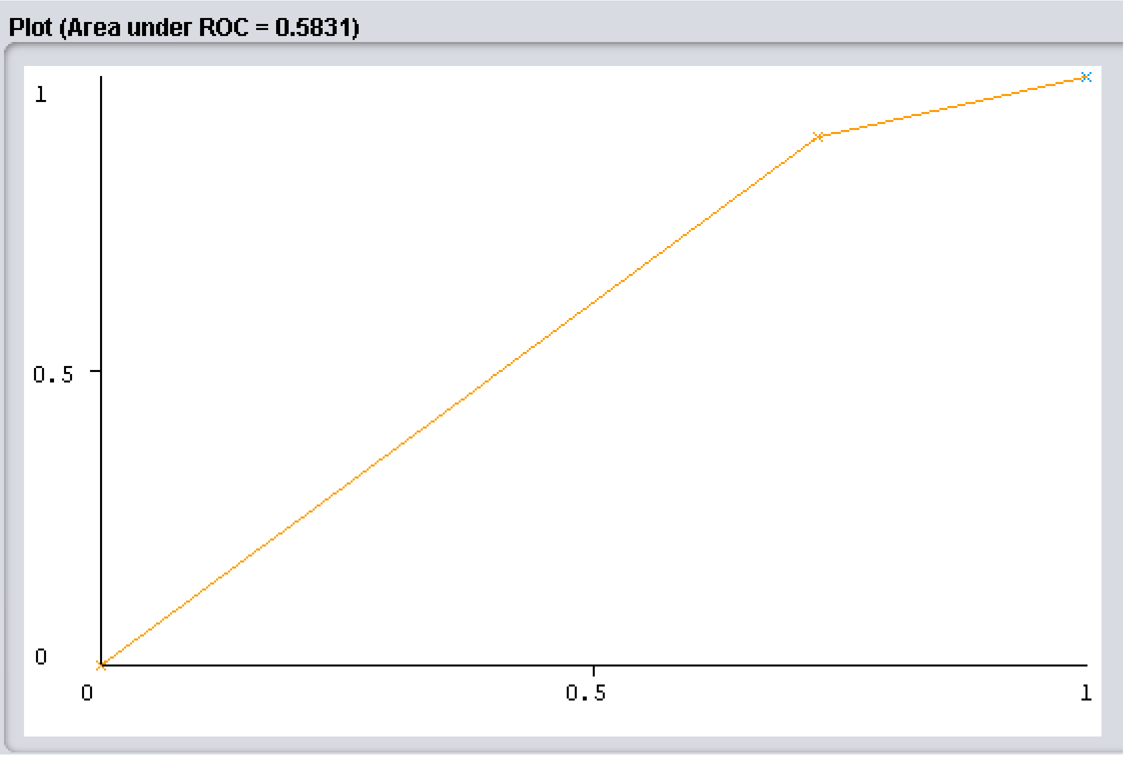


Ilustración 4: ROC BreastCancer4

- **Configuration Normalized Poly Kernel 2**

  Exponencial = 10

TP Rate	*FP Rate*	*Precision*	*Recall*	*F-Measure*	*MCC*	*ROC*
0,970	0,953	0,707	0,970	0,818	0,043	0,509

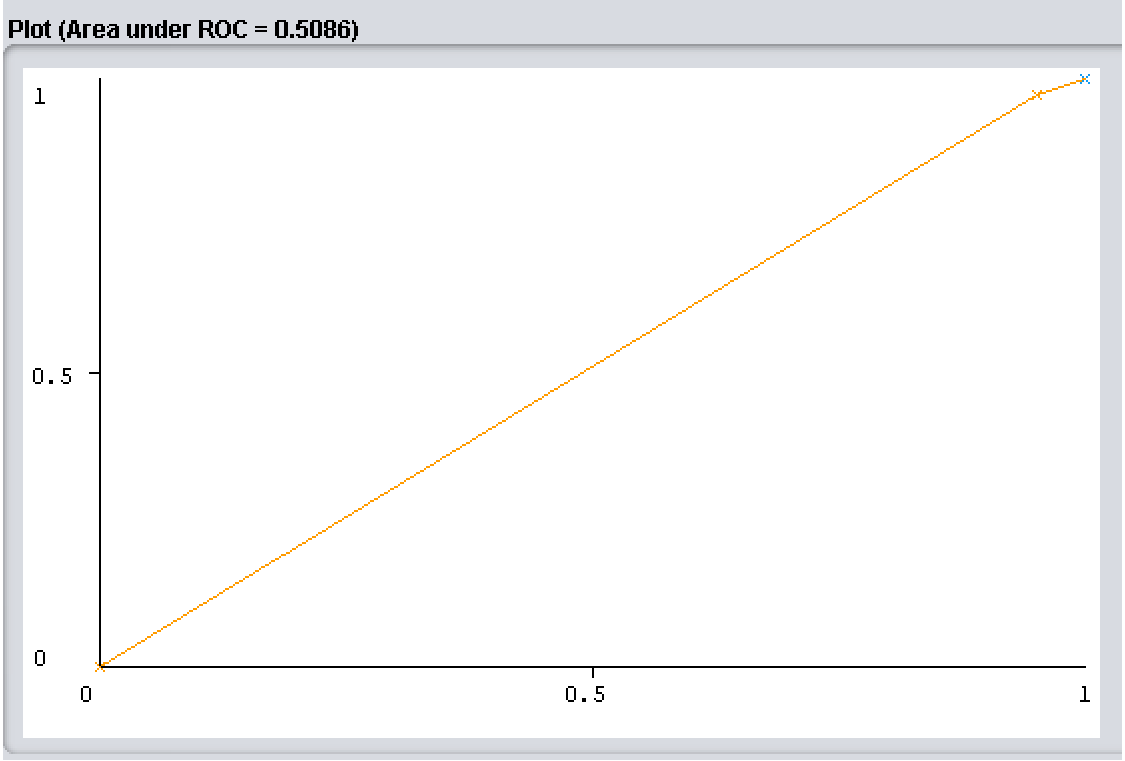


Ilustración 5: ROC BreastCancer5

- **Configuration Normalized Poly Kernel 3**

  Exponencial = 100

TP Rate	*FP Rate*	*Precision*	*Recall*	*F-Measure*	*MCC*	*ROC*
0,970	0,953	0,707	0,970	0,818	0,043	0,509

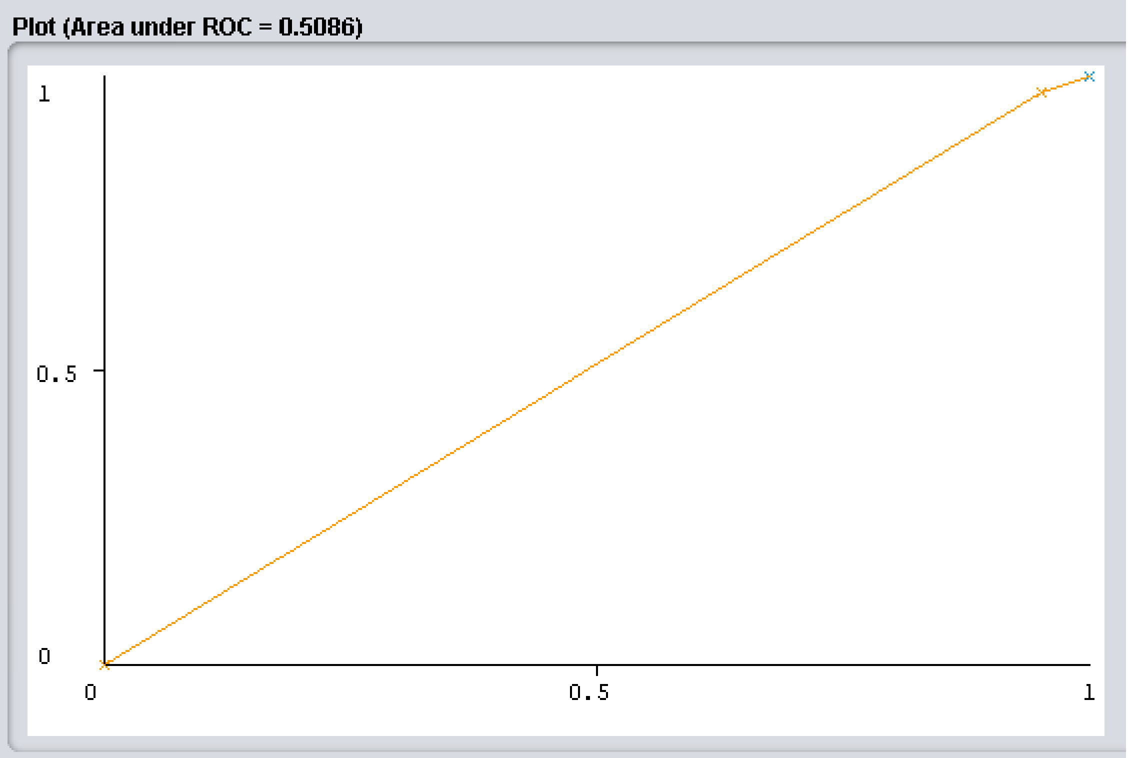

Plot (Area under ROC = 0.5086)

Ilustración 6: ROC BreastCancer6

- **Configuration RBF Kernel 1**

  Gamma = 0.01

TP Rate	*FP Rate*	*Precision*	*Recall*	*F-Measure*	*MCC*	*ROC*
1,000	1,000	0,703	1,000	0,825	?	0,500

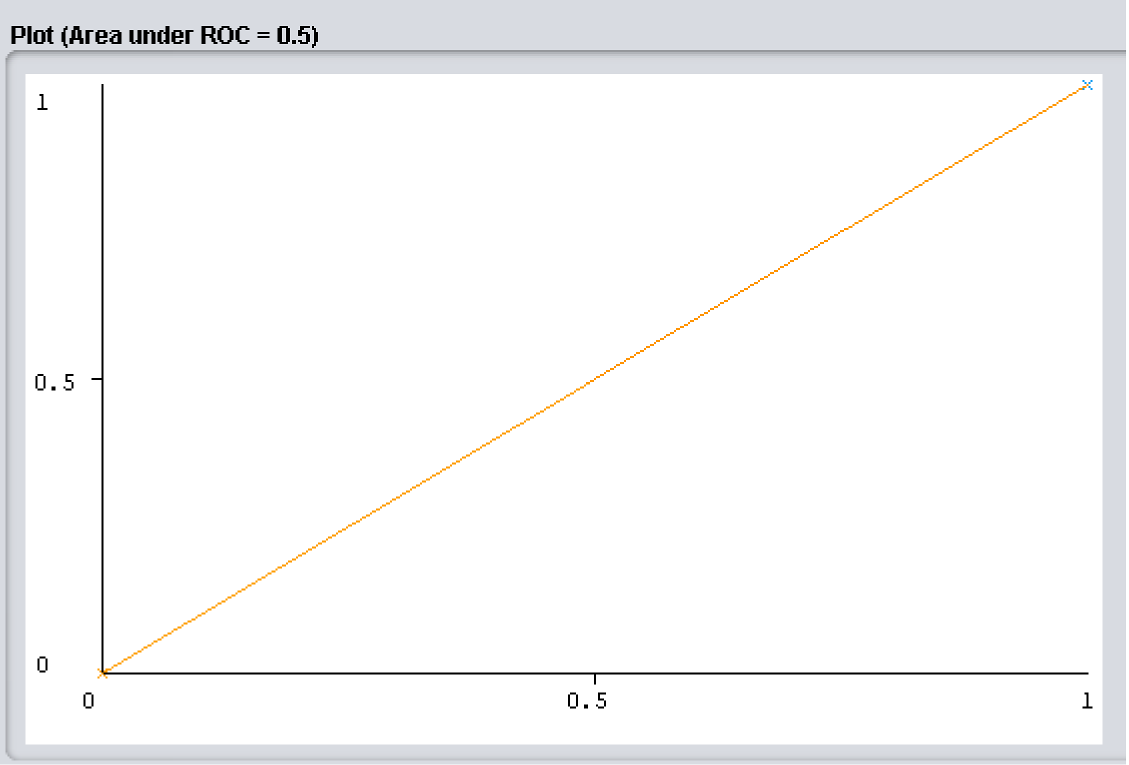


Ilustración 7: ROC BreastCancer7

- **Configuration RBF Kernel 2**

  Gamma = 0.1

TP Rate	*FP Rate*	*Precision*	*Recall*	*F-Measure*	*MCC*	*ROC*
0,950	0,741	0,752	0,950	0,840	0,303	0,605

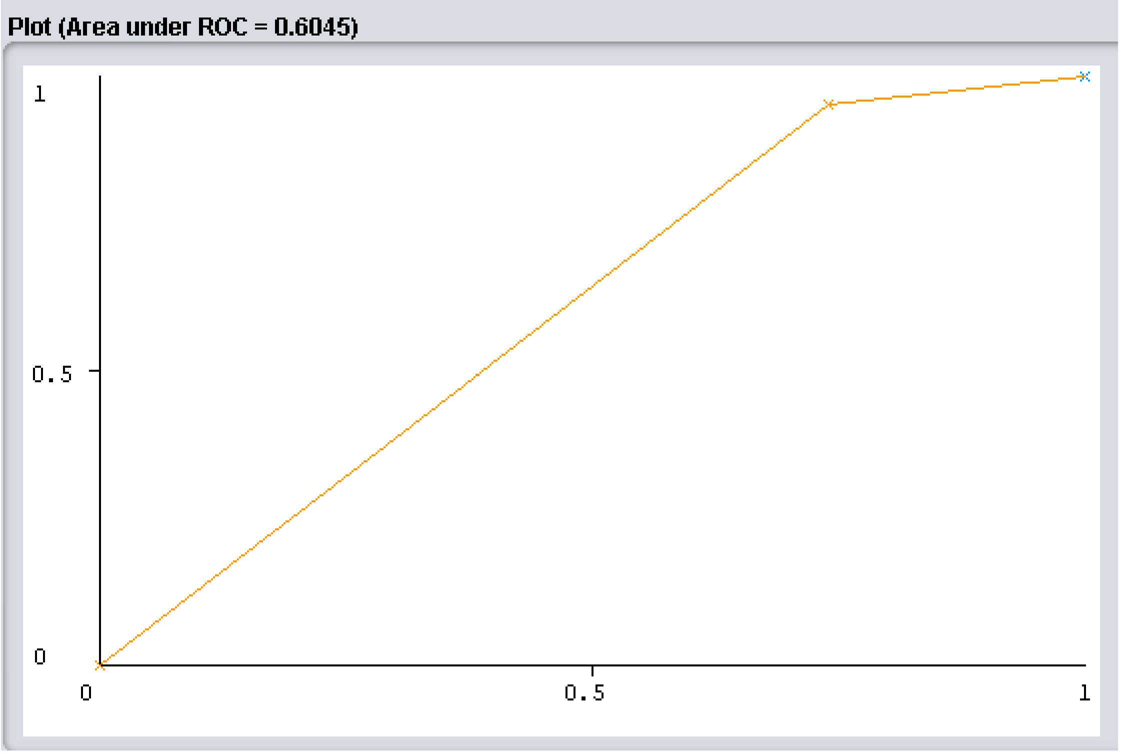


Ilustración 8: ROC BreastCancer8

- **Configuration RBF Kernel 3**

  Gamma = 1

TP Rate	*FP Rate*	*Precision*	*Recall*	*F-Measure*	*MCC*	*ROC*
0,970	0,953	0,707	0,970	0,818	0,043	0,509

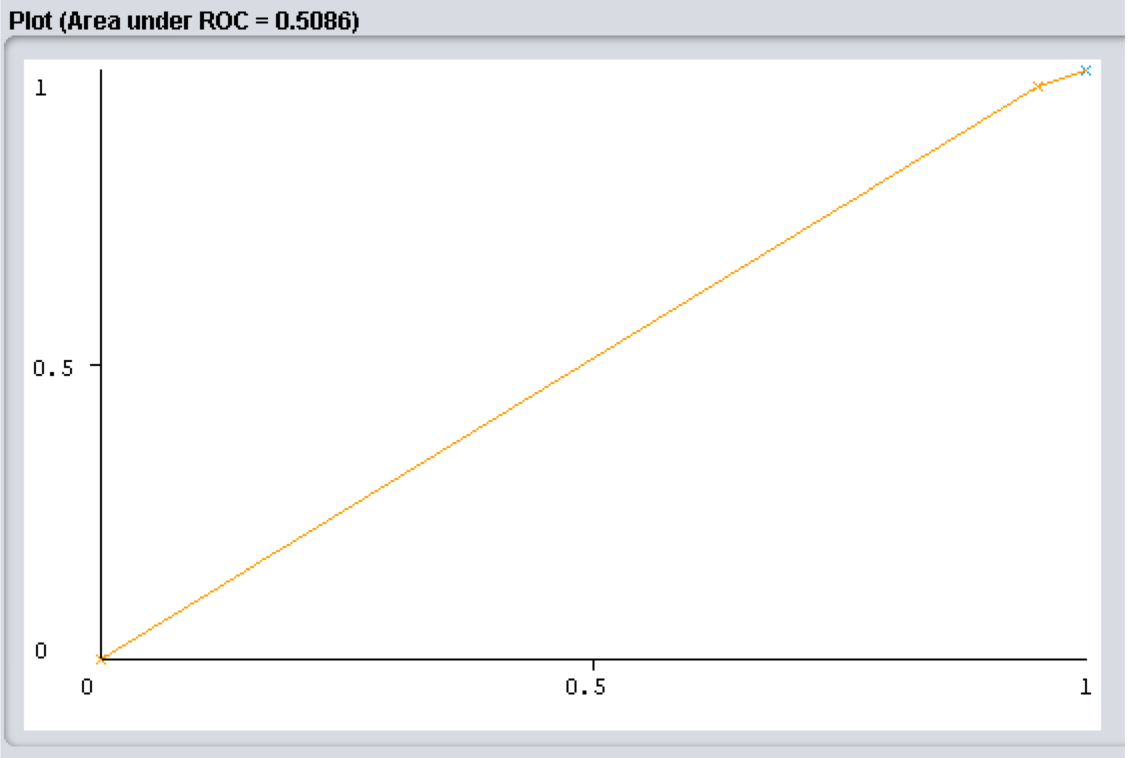


Ilustración 9: ROC BreastCancer9

First of all we base our choices on the function type of our kernel. In the case of polynomial kernels we have tested with a polynomial of different degrees marked by the exponent parameter and in the case of the RBF kernel we modify gamma, which is a parameter that modifies its amplitude. We have been giving ascending values in the three kernels to check their effect.

Regarding the results, we can verify that polynomials of a higher degree or with a greater kernel amplitude do not imply better performance, but in this case the configurations with lower exponent and gamma values have given better results. However, in values such as Precision, this relationship is not followed.

HEPATITIS DATASET

➔Introduction

The main idea of this classification problem is given a dataset about patients which suffer from hepatitis and they a series of characteristics which determine if they died or live because train a SMO classification model and evaluate its performance.

In this dataset we have 155 instances in total of which 129 of one class and 26 instances of another class. Each instance has 20 attributes, one of them is the class attribute. Because of we have only two classes we are faced with a binary classification problem.

➔Number of classes

In this case our class attribute can take only ***two values***, *die* and *live*. Because of that we are faced to a ***binary classification problem***.

➔Number of attributes

In this case we have **twenty attributes** which are *age, sex, steroid, antivirals, fatigue, malaise, anorexia, liver big, liver firm, spleen palpable, spiders, ascites, varices, bilirubin, alk phosphate, sgot, albumin, protime, histology* and *class.*

➔Number of samples

In this case in particular the number of samples or instances is **155.**

➔ Performance metrics values by configuration

➔ Configuration PolyKernel 1

Exponencial = 1

TP Rate	*FP Rate*	*Precision*	*Recall*	*F-Measure*	*MCC*	*ROC*
0,625	0,081	0,667	0,625	0,645	0,557	0,772

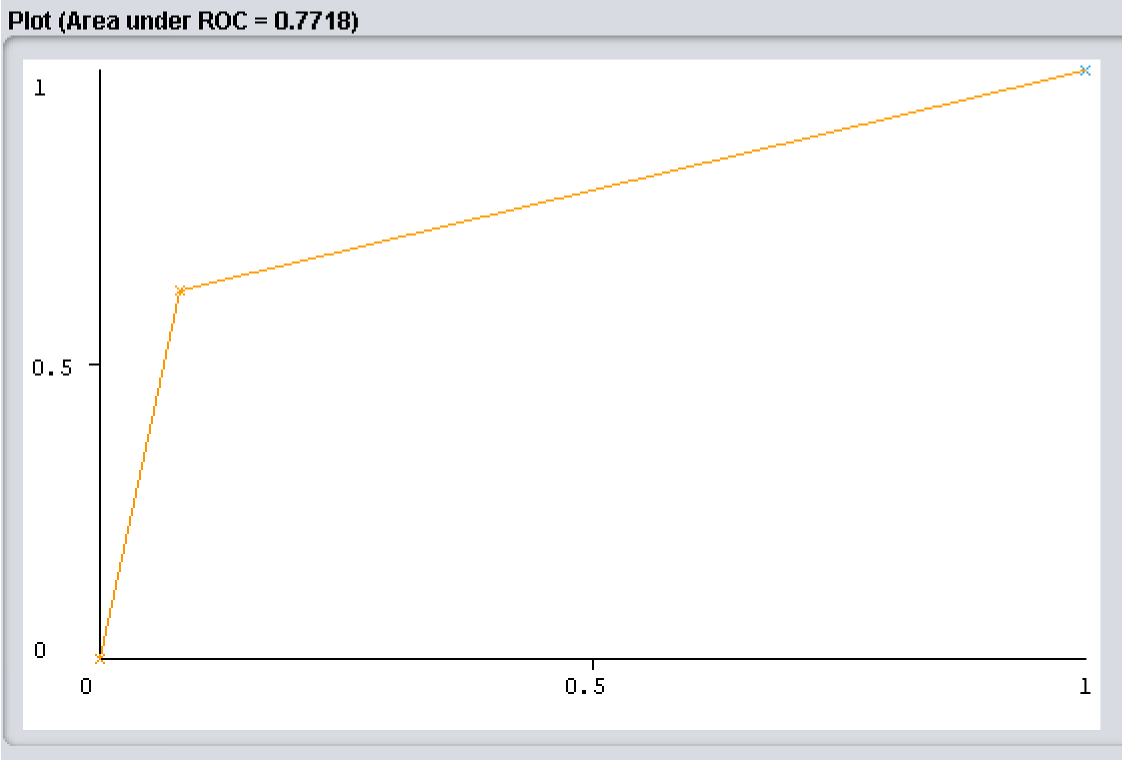
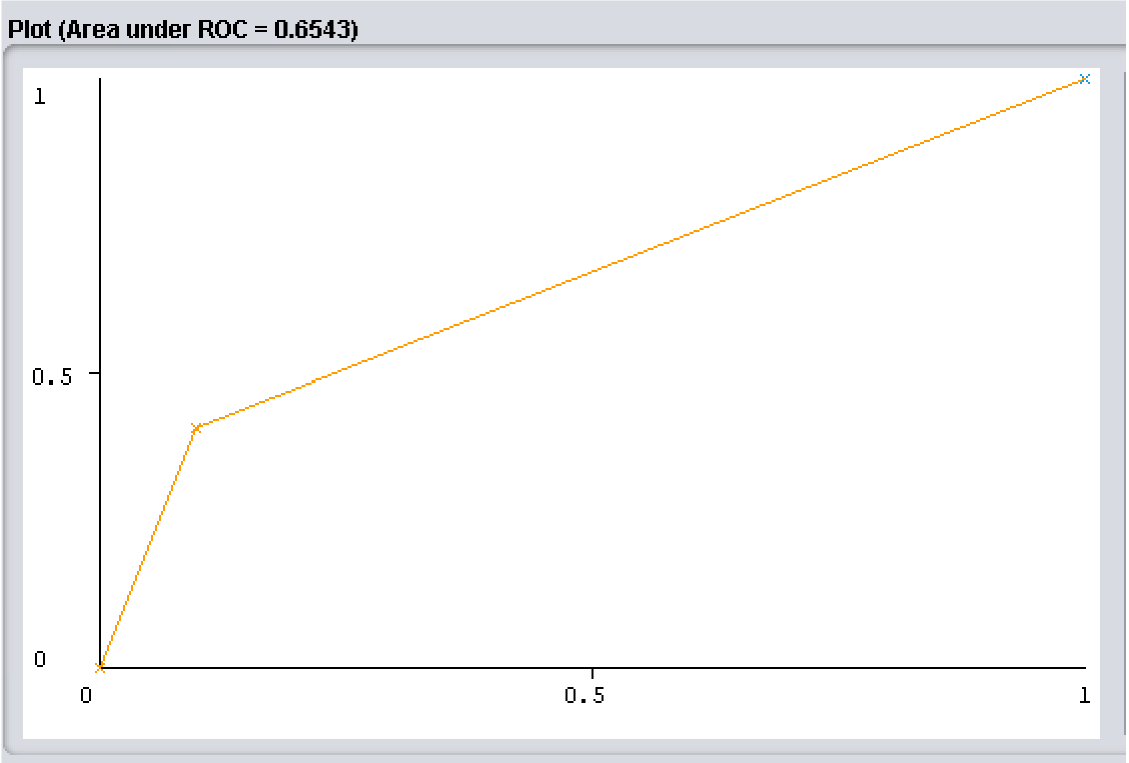


Ilustración 10: ROC Hepatitis1

➔ Configuration PolyKernel 2

Exponencial = 10

TP Rate	*FP Rate*	*Precision*	*Recall*	*F-Measure*	*MCC*	*ROC*
0,406	0,098	0,520	0,406	0,456	0,340	0,654


Plot (Area under ROC = 0.6543)

- **Configuration PolyKernel 3**

Exponencial = 100

TP Rate	*FP Rate*	*Precision*	*Recall*	*F-Measure*	*MCC*	*ROC*
1,000	1,000	0,206	1,000	0,342	?	0,500

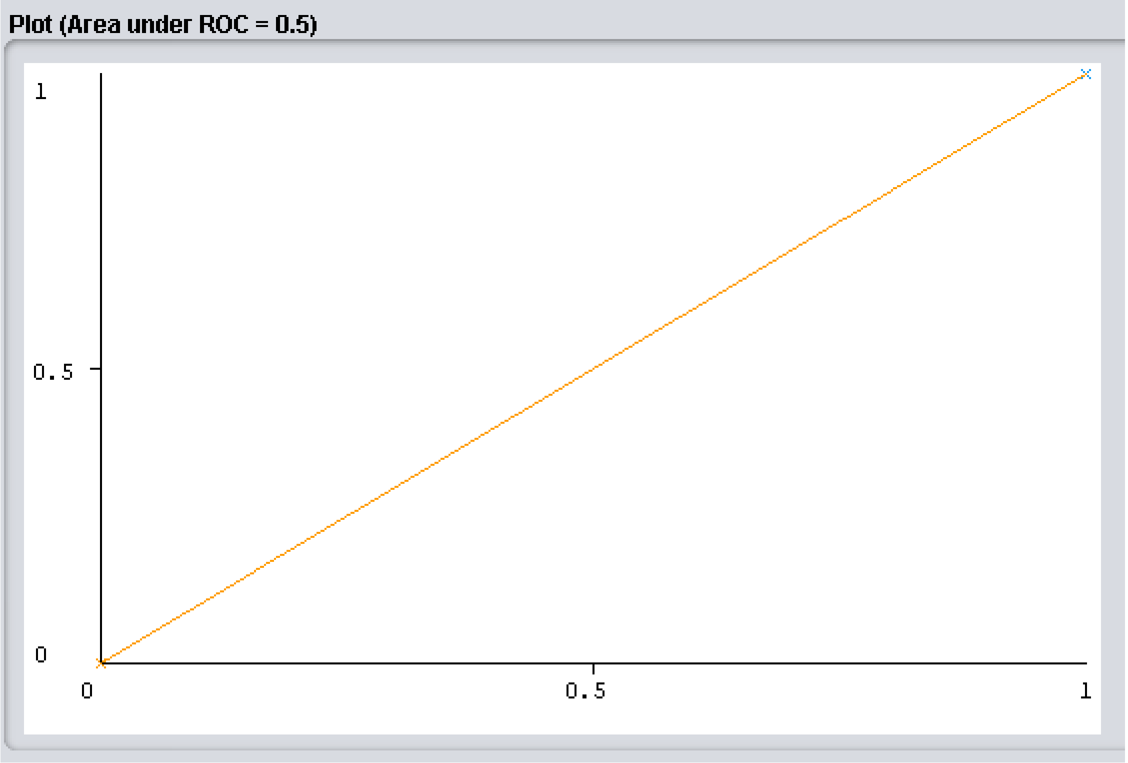


Ilustración 11: ROC Hepatitis3

- **Configuration Normalized Poly Kernel 1**

Exponencial = 1

TP Rate	*FP Rate*	*Precision*	*Recall*	*F-Measure*	*MCC*	*ROC*
0,344	0,057	0,611	0,344	0,440	0,362	0,643

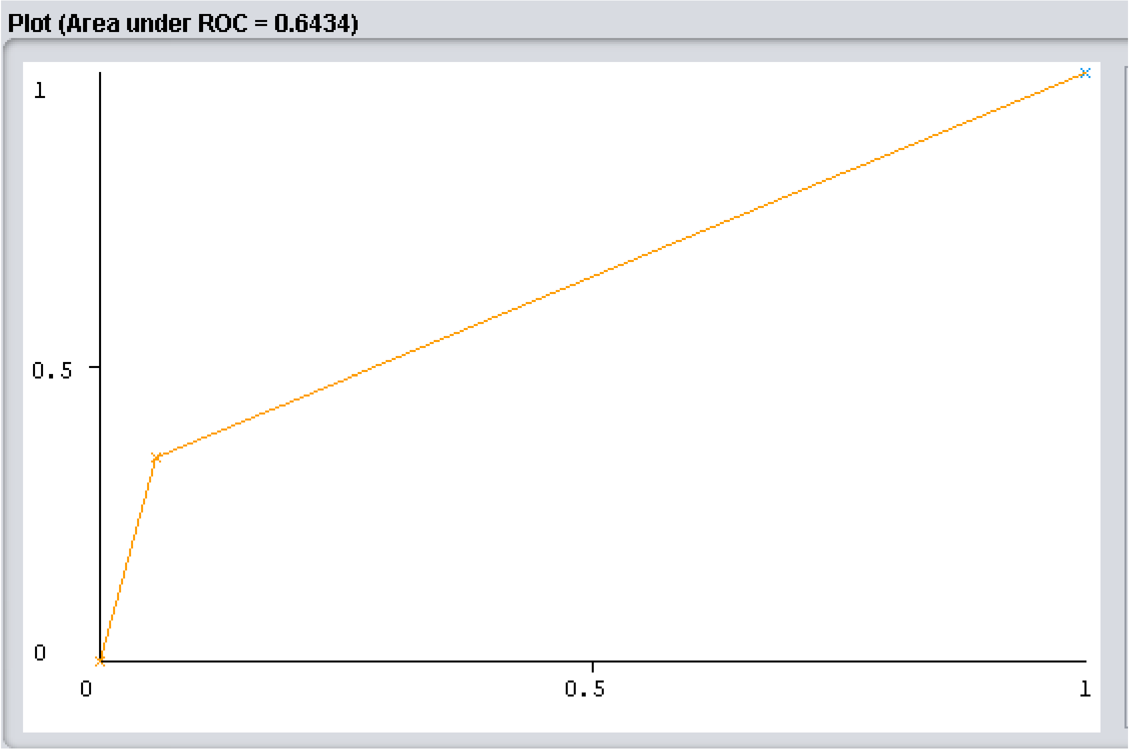
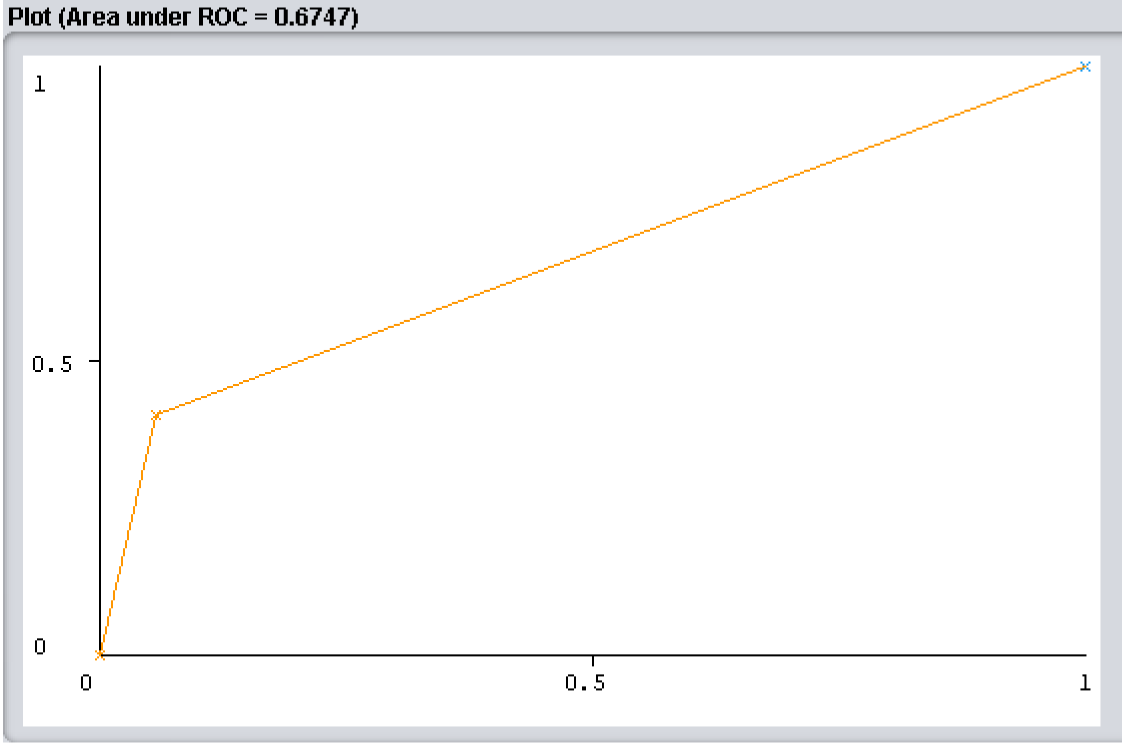


Ilustración 12: ROC Hepatitis4

➔ Configuration Normalized Poly Kernel 2

Exponencial = 10

TP Rate	*FP Rate*	*Precision*	*Recall*	*F-Measure*	*MCC*	*ROC*
0,406	0,057	0,650	0,406	0,500	0,422	0,675

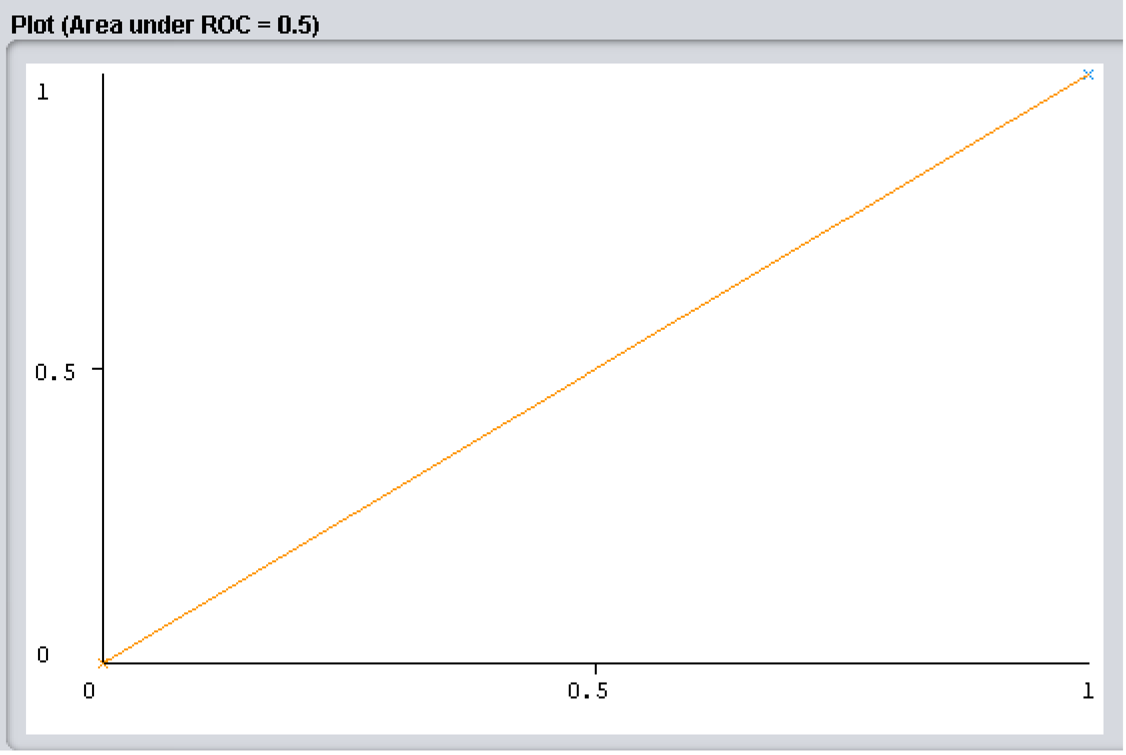


Plot (Area under ROC = 0.6747)

- **Configuration Normalized Poly Kernel 3**

Exponencial = 100

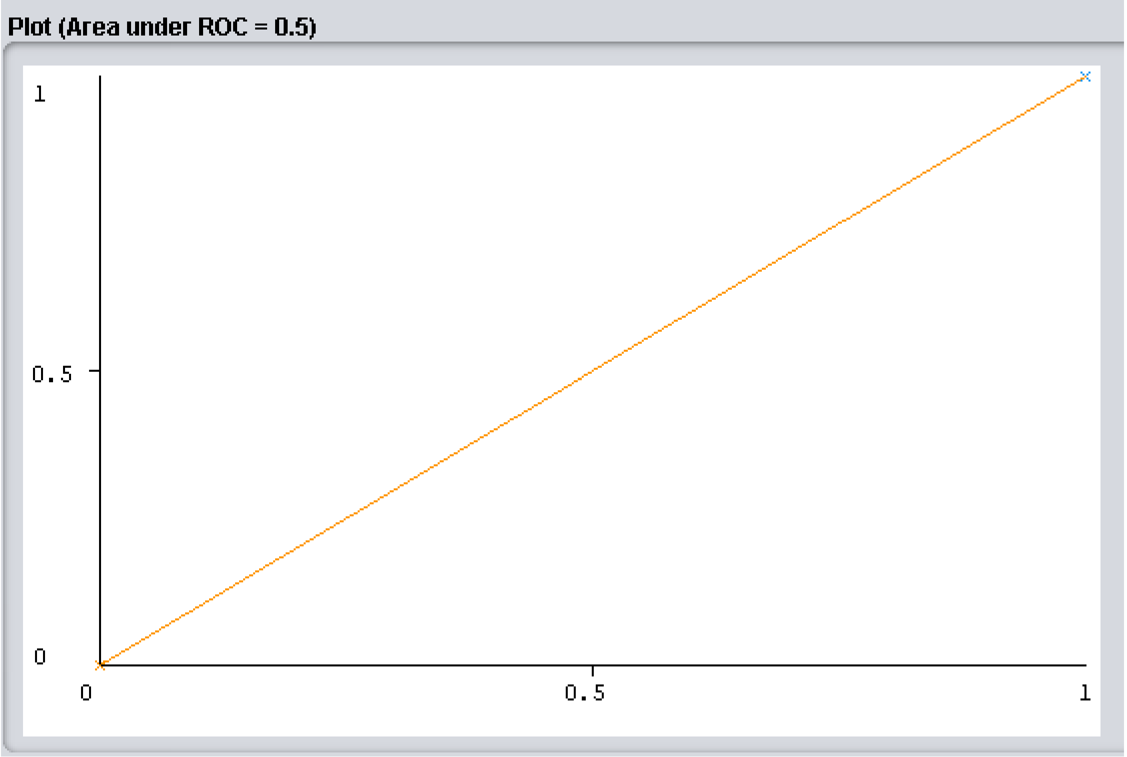
TP Rate	*FP Rate*	*Precision*	*Recall*	*F-Measure*	*MCC*	*ROC*
0,000	0,000	?	0,000	?	?	0,500



➔ Configuration RBF Kernel 1

Gamma = 0,01

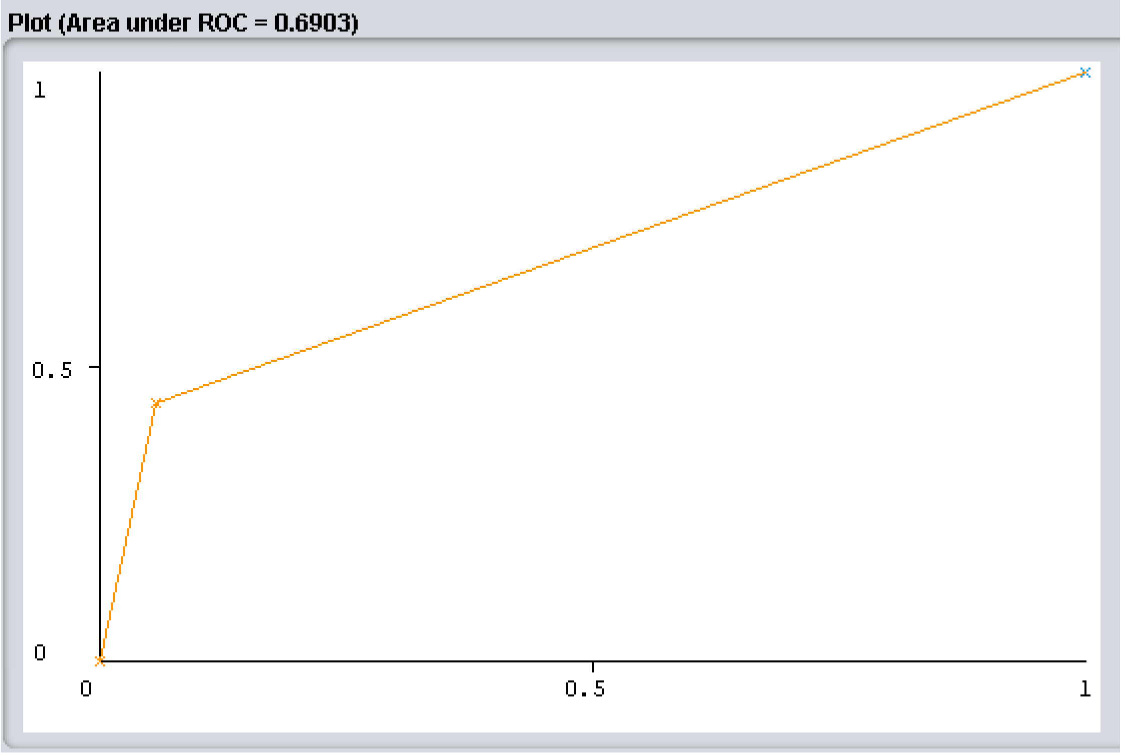
TP Rate	*FP Rate*	*Precision*	*Recall*	*F-Measure*	*MCC*	*ROC*
0,000	0,000	?	0,000	?	?	0,500


Plot (Area under ROC = 0.5)

➔ Configuration RBF Kernel 2

Gamma = 0,1

TP Rate	*FP Rate*	*Precision*	*Recall*	*F-Measure*	*MCC*	*ROC*
0,438	0,057	0,667	0,438	0,528	0,450	0,690

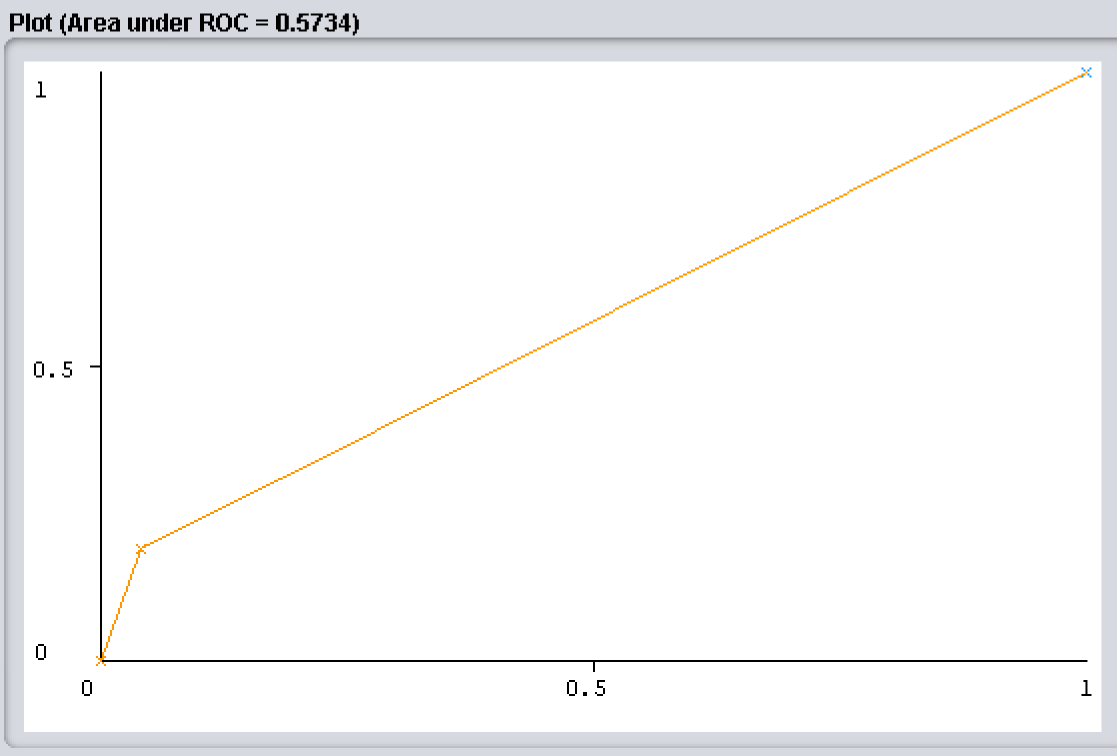


Plot (Area under ROC = 0.6903)

- ➔ **Configuration RBF Kernel 3**

  Gamma = 1

TP Rate	*FP Rate*	*Precision*	*Recall*	*F-Measure*	*MCC*	*ROC*
0,188	0,041	0,545	0,188	0,279	0,231	0,573



In this data set, something similar to the previous one occurs, small values of exponent and gamma give better results compared to the larger ones except in RBF adding that in this case the intermediate values of normalized poly kernel and RBF give better results than the small ones with a greater difference in the latter case.

POST-OPERATIVE PATIENTS DATASET

➔Introduction

The main idea of this classification problem is given a dataset about patients which have passed an operation and they are in a postoperative recovery area waiting to a decision of where they should be sent next. Depending on a number of temperature measurements we train a SMO classification model and evaluate its performance.

In this dataset we have 90 instances in total of which they are divided in three categories or classes, 2 of class I, 24 of class S and 64 of class A. Each instance has 9 attributes, one of them is the class attribute. Because of we have three possible classes we are faced with a multiclass classification problem.

➔Number of classes

In this case our class attribute can take only **three *values***, I, S and *A*. Because of that we are faced to a **multiclass *classification problem***.

➔Number of attributes

In this case we have **nine attributes** which *l-core, l-surf, l-02, l-bp, surf-stbl, core-stbl, bp-stbl, comfort* and *class.*

➔Number of samples

In this case in particular the number of samples or instances is **90.**

- **Performance metrics values by configuration**

  - **Configuration PolyKernel 1**

    Exponencial = 1

TP Rate	*FP Rate*	*Precision*	*Recall*	*F-Measure*	*MCC*	*ROC*
0,678	0,723	?	0,678	?	?	0,4777

  - **Configuration PolyKernel 2**

    Exponencial = 10

TP Rate	*FP Rate*	*Precision*	*Recall*	*F-Measure*	*MCC*	*ROC*
0,600	0,682	?	0,600	?	?	0,459

  - **Configuration PolyKernel 3**

    Exponencial = 100

TP Rate	*FP Rate*	*Precision*	*Recall*	*F-Measure*	*MCC*	*ROC*
0,711	0,711	?	0,711	?	?	0,500

➔ **Configuration Normalized Poly Kernel 1**

Exponencial = 1

TP Rate	*FP Rate*	*Precision*	*Recall*	*F-Measure*	*MCC*	*ROC*
0,711	0,711	?	0,711	?	?	0,50

➔ **Configuration Normalized Poly Kernel 2**

Exponencial = 10

TP Rate	*FP Rate*	*Precision*	*Recall*	*F-Measure*	*MCC*	*ROC*
0,678	0,723	?	0,678	?	?	0,477

➔ **Configuration Normalized Poly Kernel 3**

Exponencial = 100

TP Rate	*FP Rate*	*Precision*	*Recall*	*F-Measure*	*MCC*	*ROC*
0,667	0,727	?	0,667	?	?	0,470

➔ **Configuration RBF Kernel 1**

Gamma = 0,01

TP Rate	*FP Rate*	*Precision*	*Recall*	*F-Measure*	*MCC*	*ROC*
0,667	0,727	?	0,667	?	?	0,470

➔ **Configuration RBF Kernel 2**

Gamma = 0,1

TP Rate	*FP Rate*	*Precision*	*Recall*	*F-Measure*	*MCC*	*ROC*
0,711	0,711	?	0,711	?	?	0,500

➔ **Configuration RBF Kernel 3**

Gamma = 1

TP Rate	*FP Rate*	*Precision*	*Recall*	*F-Measure*	*MCC*	*ROC*
0,678	0,723	?	0,678	?	?	0,477

In this last dataset there is not a uniform order between the different kernels since with a polynomial kernel the highest value of the best result but the lowest is better than the intermediate and in the case of normalized polynomials the intermediate value gives the best result followed by the highest and finally the minor. In the case of RBF, the intermediate is the one that gives the best result followed by the highest. As we can see, none of the three configurations follows an order in relation to the modified values, making it difficult to choose.

➔ **Evaluation of the results**

DatasetBreast Cancer	TP rate	FP rate	Precision	Recall	F-mesure	ROC
PolyKernel1	0,866	0,671	0,753	0,866	0,806	0,598
PolyKernel10	0,547	0,376	0,775	0,547	0,641	0,585
PolyKernel100	1,00	1,00	0,703	1,00	0,825	0,500
Normalized1	0,896	0,729	0,744	0,896	0,813	0,583
Normalized10	0,970	0,953	0,707	0,970	0,818	0,509
Normalized100	0,970	0,953	0,707	0,970	0,818	0,509
RBF0.01	1,00	1,00	0,703	1,00	0,825	0,500
RBF0.1	0,950	0,741	0,752	0,950	0,840	0,605
RBF1	0,970	0,953	0,707	0,970	0,818	0,509

Dataset Hepatitis	TP rate	FP rate	Precision	Recall	Fmeasure	ROC
PolyKernel1	0,625	0,081	0,667	0,625	0,645	0,772
PolyKernel10	0,406	0,098	0,520	0,406	0,456	0,654
PolyKernel100	1,00	1,00	0,206	1,00	0,342	0,500
Normalized1	0,344	0,057	0,611	0,344	0,440	0,643
Normalized10	0,406	0,057	0,650	0,406	0,500	0,675
Normalized100	0,00	0,00	?	?	?	0,500
RBF0.01	0,00	0,00	?	?	?	0,500
RBF0.1	0,438	0,057	0,667	0,438	0,528	0,690
RBF1	0,188	0,041	0,545	0,188	0,279	0,573

Dataset PostOperative	TP rate	FP rate	Precision	Recall	Fmeasure	ROC
PolyKernel1	0,678	0,723	?	0,678	?	0,477
PolyKernel10	0,600	0,682	?	0,600	?	0,459
PolyKernel100	0,711	0,711	?	0,711	?	0,500
Normalized1	0,711	0,711	?	0,711	?	0,500
Normalized10	0,678	0,723	?	0,678	?	0,477
Normalized100	0,667	0,727	?	0,667	?	0,470
RBF0.01	0,667	0,727	?	0,667	?	0,470
RBF0.1	0,711	0,711	?	0,711	?	0,500
RBF1	0,678	0,723	?	0,678	?	0,477

➔ **Differences between datasets**

Dataset	NumSamples	ClassType	ClassDistribution	Atributte Characteristics
BreastCancer	286	Binary	201/85	Linear/Nominal
Hepatitis	155	Binary	32/123	Nominal/ Numerical
Post-Operative	70	Multiclass	2/24/64	Nominal/ Numerical

- **Last conclusions**

As a conclusion we are going to compare the results obtained in the different datasets. Looking at the performance measures we can say that Hepatitis is the dataset that best fits with an area ROC result of 0.772 in a first degree polynomial configuration, that is, a linear kernel, followed by BreastCancer with an area ROC result of 0.605 with a kernel radial and gamma 0.1 finally we have PostOperative with a better ROC result of 0.5 obtained from various configurations.

With the experience obtained we can conclude that the choice and study of the type of kernel transformation as the modifiable parameters in these determine the result and that depending on the type of data evaluated it will be more convenient to use some combinations or others without following an exact relationship.